Spiess, Keller:

A mixed approach and a distribution free multiple imputation technique for the estimation of multivariate probit models with missing values

Sonderforschungsbereich 386, Paper 102 (1998)

A mixed approach and a distribution free multiple imputation technique for the estimation of multivariate probit models with missing values

Martin Spiess* and Ferdinand Keller†

Abstract

In the present paper a mixed generalized estimating/pseudo-score equations (GEPSE) approach together with a distribution free multiple imputation technique is proposed for the estimation of regression and correlation structure parameters of multivariate probit models with missing values for an ordered categorical time invariant variable. Furthermore, a generalization of the squared trace correlation (R_T^2) for multivariate probit models, denoted as pseudo R_T^2, is proposed. A simulation study was conducted, simulating a probit model with an equicorrelation structure in the errors of an underlying regression model and using two different missing mechanisms. For a low 'true' correlation the difference between the GEPSE, a generalized estimating equations (GEE) and a maximum likelihood (ML) estimator were negligible. For a high 'true' correlation the GEPSE estimator turned out to be more efficient than the GEE and very efficient relative to the ML estimator. Furthermore, the pseudo R_T^2 was close to R_T^2 of the underlying linear model. The mixed approach is illustrated using a psychiatric data set of depressive inpatients. The results of this analysis suggest, that the depression score at discharge from a psychiatric hospital and the occurence of stressful life events seem to increase the probability of having an episode of major depression within a one-year interval after discharge. Furthermore, the correlation structure points to short-time effects on having or not having a depressive episode, not accounted for in the systematic part of the regression model.

Key words: Multivariate probit model; Panel data; Generalized estimating equations; Pseudo score equations; Multiple imputation; Pseudo R_T^2

*Department of Statistics, University of Regensburg, Germany. M. Spiess is supported by the Deutsche Forschungsgemeinschaft, SFB 386

†Department of Psychiatry, University of Ulm, Ravensburg, Germany

1 Introduction

During the past several years a large amount of work has been devoted to the problem of estimating multivariate probit models. However, unless simplifying assumptions can be made, maximum likelihood (ML) estimation of these models is hampered by the computational intractability of high-dimensional integrals.

One possibility to avoid the integration over many dimensions is to model the dependencies between the responses via random effects, the number of which determining the dimensionality of the problem. The ML estimator may then approximately be calculated using Gauss-Hermite quadratur. For clustered and/or panel probit models with one or two random effects and binary responses, see e.g. Anderson and Aitkin (1985), Bock and Aitkin (1981), Bock and Lieberman (1970), Butler and Moffit (1982) or Im and Gianola (1988). However, ML estimation of random effects models is not restricted to probit models, binary responses or the assumption of only one or two random effects. For more general models see e.g. Bock and Gibbons (1996), Conaway (1989), Gibbons, Hedeker, Charles and Frisch (1994) or Hedeker and Gibbons (1994). Other approaches for the approximate ML or non-ML estimation of random effects models are proposed e.g. by Stiratelli, Laird and Ware (1984) or Wong and Mason (1985) and in the context of generalized mixed models e.g. by Breslow and Clayton (1993), Lee and Nelder (1996), McGilchrist (1994) or Schall (1991).

Alternative non-ML approaches for the estimation of general clustered and/or panel logit or probit models have been proposed e.g. by Avery, Hansen and Hotz (1983), Gourieroux, Monfort, and Trognon (1984), Liang and Zeger (1986) or Schepers, Arminger and Küsters (1991). A survey of methods for the estimation of clustered and/or panel models with emphasis on logit models and binary responses is given e.g. by Pendergast et al. (1996).

An easy to implement and computational efficient method is the 'generalized estimating equations' (GEE) approach proposed by Liang and Zeger (1986). This approach allows the consistent estimation of regression parameters even if the correlation structure of the outcomes is misspecified (Liang & Zeger, 1986), using generalized estimating equations for the estimation of regression parameters and simple functions of residuals for the estimation of the correlation structure parameters. If the correlation structure is correctly specified, then the loss of efficiency of the regression parameter estimators is small relative to the ML regression parameter estimators. On the other hand, the parameter estimators modeling the correlation structure may be very inefficient (Liang, Zeger & Qaqish, 1992). For a more efficient estimation of both types of parameters Prentice (1988) proposed the estimation of both sets of parameters by generalized estimating equations.

Modeling the correlation structure of the observable outcomes, the GEE approach and its extensions described so far originally were not intended for the estimation of functions of correlations of not observable, continuous response variables, given the covariates. Starting with a threshold model in the context of probit models, however, the concept of models with partially observable (e.g. binary) responses is used in many contexts (e.g. Ashford & Sowden, 1970; Heckman, 1981; Muthén, 1984; Pearson, 1900; Schepers et al., 1991). In these cases not only the regression parameters but also functions of the correlations of the latent responses given the covariates, i.e. the underlying correlations, are of interest. Therefore, based upon the extended GEE approach proposed by Prentice (1988), in their work Qu, Williams, Beck and Medendorp (1992) and Qu, Piedmonte and Williams (1994) propose the simultaneous estimation of both sets of parameters, i.e. regression parameters and functions of the underlying correlations, henceforth called correlation structure parameters, using generalized estimating equations.

In the present paper a different approach for the simultaneous estimation of regression and correlation structure parameters is proposed. In contrast to the approach proposed by Qu et al. (1992) and Qu et al. (1994), the correlation structure parameters are estimated using pseudo-score equations. Since the regression parameters are estimated using generalized estimating equations, this mixed approach will be called GEPSE approach (generalized estimating/pseudo score equations approach). Both sets of parameters are calculated as if they were orthogonal, thereby preserving the robustness of the regression parameter estimators with respect to misspecification of the correlation matrix. The use of generalized estimating equations for the estimation of regression parameters was shown to lead to more efficient estimators than using a three-stage approach as proposed e.g. by Schepers et al. (1991) in a Monte Carlo experiment (Spiess & Hamerle, 1995). In contrast to the pseudo-ML approach proposed by Gourieroux et al. (1984) where the regression parameters are estimated under the assumption of independence, using the mixed approach, the regression parameters are estimated taking the associations between the responses into account. Although the mixed approach can be used to estimate general multivariate probit models, the present paper focuses on the estimation of cluster or panel models with binary responses.

The proposed approach will be illustrated analysing the impact of 'stressful life events', 'depression score at discharge', 'age' and 'gender' of patients as well as time effects upon the probability of having a depressive episode within a one year interval after discharge from a psychiatric state hospital. Two types of correlation structures in the assumed latent 'depressivity' given the covariates are considered: Equicorrelation and an autocorrelation-like structure. An equicorrelation structure for example could point to a prevailing impact of individual specific factors not accounted for in the systematic part of the model, maintaining the 'depressivity' level over time. On the other hand, if an autocorrelation-like structure is present, then the assumption of a prevailing impact of factors with decreasing effects over time on the 'depressivity' level, again given the covariates, would be plausible.

Unfortunately, the 'depression score at discharge' is not observed for all patients. The problem of missing data is a common problem in many applications.

However, recent advances have led to a wide variety of strategies for coping with this problem in statistical inference (e.g. Little & Rubin, 1987; Little, 1992). One popular method is the method of multiple imputation (e.g. Rubin, 1987), where several complete data sets are created filling in the missing values. The advantage of this method is that standard methods for the analysis of complete data sets can be applied, i.e. this technique is not tied to one particular estimation method. In the present paper a distribution free approach is used, which is based upon a regression of the variable which is not observed for all patients on all other variables for the complete cases and the imputation of predicted values for incomplete cases (e.g. Heitjan & Little, 1991, or, Schenker & Taylor, 1996). However, since the variable 'depression score at discharge' is treated as an ordered categorical variable, instead of using a linear regression, the variate 'ranks of the depression score at discharge' is regressed on all other variables.

To assess the fit of the systematic part of the model used to analyse the data set described above, the pseudo R^2 proposed by McKelvey and Zavoina (1975) for ordinal probit models with uncorrelated responses is extended to a pseudo R^2_T for multivariate models which is a generalization of the trace correlation in multivariate linear regression models (Hooper, 1959).

This article is organized as follows. In section 2 the model is described and the notation introduced. In section 3 the mixed estimation procedure and the asymptotic properties of the estimator are presented. A sketch of the proof of the asymptotic properties is given in the Appendix. Section 4 describes the distribution free multiple imputation technique. Section 5 provides the results of a simulation study, comparing the proposed GEPSE estimator with a GEE and a ML estimator in finite samples. The mixed estimation procedure is illustrated using a psychiatric dataset in section 6. Conclusions can be found in section 7.

2 The Model

Let N $(n = 1, \ldots, N)$ be the number of clusters (e.g. subjects), T $(t = 1, \ldots, T)$ be the number of observations within every cluster and $y_n = (y_{n1}, \ldots, y_{nT})'$ the vector of observable binary responses for the nth cluster. Let $x_{nt} = (x_{nt1}, \ldots, x_{ntP})'$ denote the $(P \times 1)$ vector of covariates associated with the tth observation of the nth cluster, X_n the $(T \times P)$ matrix of covariates associated with the nth cluster and X the $(NT \times P)$ matrix having full column rank associated with all NT observations. All types of 'truly' exogenous variables are allowed, e.g. covariates which are invariant over clusters, invariant over observations within clusters or covariates varying over all clusters and observations.

Throughout a threshold model (Pearson, 1900)

$$y^*_{nt} = x'_{nt}\beta^\star + \upsilon_{nt} \quad \text{and} \quad y_{nt} = \begin{cases} 1 & \text{if } y^*_{nt} > 0, \\ 0 & \text{otherwise,} \end{cases}$$

is assumed, where y^*_{nt} is an unobservable, continuous response variable, $\beta^\star$ is the unknown regression parameter vector and υ_{nt} is an unobservable error term distributed independently of the covariates. For the multivariate probit model, let $\upsilon_n \sim N(0, \Sigma)$, where $\upsilon_n = (\upsilon_{n1}, \ldots, \upsilon_{nT})'$, and $\Sigma = \mathrm{V}^{1/2}\,\mathrm{R}\,\mathrm{V}^{1/2}$, where $\mathrm{V} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_T^2)$ denotes a diagonal matrix with the diagonal elements being the variances, σ_t^2, of υ_{nt}, and R is a correlation matrix with elements $\rho_{tt'}$, the pairwise correlations between observation points t and t'. Throughout, ρ denotes the vector of all $T(T-1)/2$ off diagonal elements of R, i.e. $\rho = (\rho_{21}, \rho_{31}, \rho_{32}, \ldots, \rho_{T(T-1)})'$. The structure of R depends upon the process in the error terms υ_{nt}. For example, a stationary first-order autoregressive process (AR(1) process) leads to an AR(1) structure in the corresponding correlation matrix, i.e. $\rho_{tt'} = \vartheta^{|t-t'|}$, $|\vartheta| < 1$. If the error term υ_{nt} is composed of two independent terms, one cluster specific and one observation specific, π_n and ϵ_{nt}, say, then the corresponding correlation matrix has an equicorrelation structure, i.e. $\rho_{tt'} = \vartheta$ for all t, t' $(t \neq t')$. Of course, other correlation structures could be modeled. Observations from different blocks are assumed to be independent.

In the sequel let $\Phi(\cdot)$ denote the standard normal cumulative distribution function, $\varphi(\cdot)$ the standard normal density function, $\Phi(\cdot, \cdot, \rho_{tt'})$ the standard bivariate normal cumulative distribution function and $\varphi(\cdot, \cdot, \rho_{tt'})$ the standard bivariate normal density function.

3 Estimation of complete data sets

In the model of Section 2 only the parameter vectors $\beta_t = \sigma_t^{-1}\beta^\star$ are identifiable. Therefore, the usual restriction $\sigma_t = \sigma$ for all t will be adopted[1]. The identifiable regression parameter then is $\beta = \sigma^{-1}\beta^\star$. Although in this paper only probit models with correlated binary responses are considered, the proposed approach can easily be extended to the estimation of more general probit models with ordered categorical or mixed continuous/categorical correlated responses.

The two sets of parameters, β and ϑ, where in contrast to Liang and Zeger (1986) or Prentice (1988) ϑ is a function of the underlying correlations, can be estimated using the generalized estimating equations for the regression parameters (Liang & Zeger, 1986; Prentice, 1988)

$$\sum_n \mathrm{A}_n'\, \Omega_n^{-1}\, \mathrm{e}_n = 0 \tag{1}$$

and the pseudo-score equations for the estimation of the correlation structure parameters

$$\frac{\partial \rho}{\partial \vartheta} \sum_n \mathrm{B}_n'\, \mathrm{W}_n^{-1}\, \mathrm{v}_n = 0, \tag{2}$$

[1] This constraint is more restrictive than necessary and could be relaxed in what follows.

where ρ is considered as a function $\mathrm{f}(\vartheta)$ of the structural parameter ϑ, which may be a vector or a scalar. For example, if an AR(1) structure is assumed, $\mathrm{f}(\vartheta) = (\vartheta^1, \vartheta^2, \vartheta^1, \ldots, \vartheta^{T-1}, \ldots, \vartheta^1)'$, where ϑ is a scalar. If an equicorrelation structure is assumed, then $\mathrm{f}(\vartheta) = 1_{T(T-1)/2}\vartheta$, where $1_{T(T-1)/2}$ is a $(T(T-1)/2 \times 1)$ vector with all elements equal to unity and ϑ is again a scalar.

For the probit model considered in this article,

$$\mathrm{e}_n = y_n - \Phi(X_n\beta),$$

$$\Omega_n = \mathrm{Cov}(y_n),$$

with diagonal elements $\Phi(x'_{nt}\beta)(1 - \Phi(x'_{nt}\beta))$ and covariance, i.e. off diagonal element, $\Phi(x'_{nt}\beta, x'_{nt'}\beta, \hat{\rho}_{tt'}) - \Phi(x'_{nt}\beta)\Phi(x'_{nt'}\beta)$ in the tth row and t'th column $(t \neq t')$ and

$$\mathrm{A}'_n = X'_n \mathrm{diag}(\varphi(x'_{n1}\beta), \ldots, \varphi(x'_{nT}\beta)).$$

The elements of the $(T(T-1)/2 \times 1)$ vector v_n are $(2y_{nt} - 1)(2y_{nt'} - 1)$,

$$\mathrm{W}_n = \mathrm{diag}(P_{n(2,1)}, \ldots, P_{n(T,T-1)}),$$

where $P_{n(t,t')} = \Pr(y_{nt}, y_{nt'} \,|\, x'_{nt}\hat{\beta}, x'_{nt'}\hat{\beta}, \rho_{tt'})$ is the probability of the variables y_{nt} and $y_{nt'}$ assuming specific values, given the covariates, the regression parameter and correlation, and

$$\mathrm{B}_n = \mathrm{diag}(\varphi(x'_{n2}\hat{\beta}, x'_{n1}\hat{\beta}, \rho_{21}), \ldots, \varphi(x'_{nT}\hat{\beta}, x'_{n(T-1)}\hat{\beta}, \rho_{T(T-1)})).$$

Note that (2) is just the vector of first derivatives of the pseudo-maximum likelihood functions

$$l(\vartheta) = \sum_n l_n(\vartheta) = \sum_n \sum_{\substack{t,t' \\ (t'<t)}} \log P_{n(t,t')}$$

with respect to ϑ, where $\sum_{\substack{t,t' \\ (t'<t)}}$ means summation over all probabilities $P_{n(2,1)}$, $P_{n(3,1)}, P_{n(3,2)}, \ldots, P_{n(T,T-1)}$. Note that $P_{n(t,t')}$ is also a function of β, so if necessary the function $l(\vartheta)$ will also be written as $l(\vartheta, \beta)$.

The corresponding estimators $\hat{\vartheta}$ are similar to the pseudo ML (PML) estimators described in Gourieroux et al. (1984), in that these estimators are calculated as if the $y_t y_{t'}$ were independent. However, contrary to Gourieroux et al. (1984) who used PML estimators for β calculated under the assumption of an independent probit model, in the approach proposed above the regression parameters are estimated taking into account the assumed structure of association between the responses. Similar to the approach proposed by Qu et al. (1992) and Qu et al. (1994) both types of parameters are estimated simultaneously and the regression parameters are estimated using generalized estimating equations. In contrast to their approach, however, the correlation structure parameters are estimated using pseudo-score equations. The GEE and the GEPSE approach will be compared with respect to efficiency in section 5 in a simulation study.

The vector of estimates $\hat{\theta} = (\hat{\beta}', \hat{\vartheta}')'$ is iteratively calculated with updated value in the $(j+1)$th iteration given by

$$\hat{\theta}_{j+1} = \hat{\theta}_j - \begin{pmatrix} -\left(\sum_n \mathrm{A}_n' \, \Omega_n^{-1} \mathrm{A}_n\right)^{-1} & 0 \\ 0 & \mathrm{D}^{-1} \end{pmatrix}_{\theta=\hat{\theta}_j} \begin{pmatrix} \sum_n \mathrm{A}_n' \, \Omega_n^{-1} \mathrm{e}_n \\ \frac{\partial \rho}{\partial \vartheta} \sum_n \mathrm{B}_n' \, \mathrm{W}_n^{-1} \, \mathrm{v}_n \end{pmatrix}_{\theta=\hat{\theta}_j}$$

where

$$\mathrm{D} = \sum_n \frac{\partial^2 \, l_n(\vartheta)}{\partial \, \vartheta \, \partial \, \vartheta'}.$$

It can be shown that $\sqrt{N}(\hat{\theta} - \theta_0)$, where θ_0 is the true value, is asymptotically normally distributed with zero mean and an asymptotic covariance matrix consistenly estimated by

$$\widehat{\mathrm{Cov}}(\hat{\theta}) = N \begin{pmatrix} \mathrm{L} & 0 \\ \mathrm{M} & \mathrm{Q} \end{pmatrix}^{-1} \begin{pmatrix} \Lambda_{11} & \Lambda_{21}' \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \begin{pmatrix} \mathrm{L} & \mathrm{M}' \\ 0 & \mathrm{Q} \end{pmatrix}^{-1},$$

where

$$\mathrm{L} = \left(- \sum_n \mathrm{A}_n' \Omega_n^{-1} \mathrm{A}_n \right)_{\theta=\hat{\theta}}, \qquad \mathrm{M} = \left(\sum_n \frac{\partial^2 \, l_n(\vartheta, \beta)}{\partial \, \vartheta \, \partial \, \beta'} \right)_{\theta=\hat{\theta}},$$

$$\mathrm{Q} = \left(\sum_n \frac{\partial^2 \, l_n(\vartheta)}{\partial \, \vartheta \, \partial \, \vartheta'} \right)_{\theta=\hat{\theta}}, \qquad \Lambda_{11} = \left(\sum_n \mathrm{A}_n' \Omega_n^{-1} \mathrm{e}_n \mathrm{e}_n' \Omega_n^{-1} \mathrm{A}_n \right)_{\theta=\hat{\theta}},$$

$$\Lambda_{21} = \left(\frac{\partial \, \rho}{\partial \, \vartheta} \sum_n \mathrm{B}_n' \mathrm{W}_n^{-1} \mathrm{v}_n \mathrm{e}_n' \Omega_n^{-1} \mathrm{A}_n \right)_{\theta=\hat{\theta}}$$

and

$$\Lambda_{22} = \left(\frac{\partial \, \rho}{\partial \, \vartheta} \left(\sum_n \mathrm{B}_n' \mathrm{W}_n^{-1} \mathrm{v}_n \mathrm{v}_n' \mathrm{W}_n^{-1} \mathrm{B}_n \right) \frac{\partial \, \rho}{\partial \, \vartheta'} \right)_{\theta=\hat{\theta}}$$

(see Appendix).

To assess the goodness of fit of ordinal probit models with uncorrelated responses, McKelvey and Zavoina (1975) proposed a pseudo R^2, that gives an estimate of the coefficient of determination R^2 of the underlying linear regression model. A generalization of R^2 for multivariate linear regression models is the squared trace correlation, proposed by Hooper (1959), defined as

$$\mathrm{R}_T^2 = T^{-1} \mathrm{tr}(\mathrm{I} - \mathrm{D}),$$

where $\mathrm{D} = (\sum_n y_n y_n')^{-1} (\sum_n \upsilon_n \upsilon_n')$, y_n is a $(T \times 1)$ vector of observable responses, I is the $(T \times T)$ identity matrix and trA denotes the trace of matrix A. The squared trace correlation, R_T^2, can be interpreted as the portion of the total variance of the jointly dependent variables that is 'explained' by the systematic part of the model (Hooper, 1959).

Now consider the transformed model of section 2

$$V^{-1/2} y_n^* = V^{-1/2} X_n \beta^\star + u_n,$$

where $y_n^* = (y_{n1}^*, \ldots, y_{nT}^*)'$, $u_n = V^{-1/2} \upsilon_n$ and $\mathrm{Var}(u_n) = R$. An estimate of the residual sum of squares and products (SSP) matrix is then given by $\widehat{SSP}_R = N\,\widehat{R}$. Let $\hat{y}_n = \hat{V}^{-1/2} X_n \hat{\beta}^\star$, then the fitted SSP matrix is $SSP_F = \sum_n (\hat{y}_n - \bar{\hat{y}})(\hat{y}_n - \bar{\hat{y}})'$, where $\bar{\hat{y}} = N^{-1} \sum_n \hat{y}_n$. Thus, an estimate of the total SSP matrix is obtained by $\widehat{SSP}_T = SSP_F + \widehat{SSP}_R$. The estimate $\widehat{R}_T^2$, or pseudo R_T^2, is then given by[2]

$$\widehat{R}_T^2 = T^{-1} \mathrm{tr}(I - (\widehat{SSP}_T)^{-1} \widehat{SSP}_R) = T^{-1} \mathrm{tr}((\widehat{SSP}_T)^{-1} SSP_F).$$

Since we restrict $\sigma_t = 1$ for all t, we have $V = I$ and $\hat{\beta} = \hat{\beta}^\star$. Note, that the above partitioning of the total SSP matrix is not entirely valid since the regression parameter estimator is not unbiased. However, since it is asymptotically unbiased, for large samples, the above partition holds asymptotically. In the case of uncorrelated responses, the pseudo R^2 proposed by McKelvey and Zavoina (1975) was the one that is closest to the OLS-R^2 from various pseudo R^2 considered in several simulation studies (Veall & Zimmermann, 1992; Veall & Zimmermann, 1996; Windmeijer, 1995). In section 5, $\widehat{R}_T^2$ will be compared with the squared trace correlation of the underlying multivariate linear model using simulated data sets.

4 A distribution free multiple imputation technique

Filling in, i.e. imputing missing values is a popular method if not all values of some of the variables considered are observed, since complete-data methods can be used. However, imputing just one value for each missing value (single imputation) overstates precision, i.e. systematically underestimates the uncertainty about which value to impute, typically leading to invalid tests and confidence intervals (Heitjan and Little, 1991; Rubin and Schenker, 1986; Rubin, 1987). In contrast, multiple imputation methods (Rubin, 1987; Rubin, 1996) lead to several (M) completed data sets, each of which is analysed using complete-data methods. To correctly account for the uncertainty due to missing data, in general each of the $M > 1$ sets of imputations should be drawn independently according to the

[2]Note that since $\mathrm{tr}(\hat{V}^{1/2} (\widehat{SSP}_T)^{-1} SSP_F \hat{V}^{-1/2}) = \mathrm{tr}((\widehat{SSP}_T)^{-1} \hat{V}^{1/2} \hat{V}^{-1/2} SSP_F) = \mathrm{tr}((\widehat{SSP}_T)^{-1} SSP_F)$, $\widehat{R}_T^2$ of the transformed model is identical to $\widehat{R}_T^2$ of the original model.

following general scheme (e.g. Rubin and Schenker, 1986). Given a model for the data, the parameters should be drawn from their approximative posterior distribution (given the observed data) and, given the drawn parameters, the missing values should be drawn as independently and identically distributed.

The variable 'depression score at discharge' ('DS') in the data set (see section 6) is characterized by a substantial portion of missing values. It is an ordered categorical time invariant variate which is collected only once for every patient, i.e. at discharge from the psychiatric hospital. In the following, let $z_n = \mathrm{rg}(x_n^{DS})$ be the rank assigned to the nth value of the variable DS. Furthermore, let the index 'obs' denote the cases having complete data on all variables and the index 'mis' denote those cases having missing data on the variable DS. For example, N_{obs} denotes the number of cases having complete data on all variables and $N_{\mathrm{mis}} = N - N_{\mathrm{obs}}$. Let $n_1 = 1, \ldots, N_{\mathrm{obs}}$ and $n_2 = N_{\mathrm{obs}} + 1, \ldots, N$.

The method proposed combines techniques described in Heitjan and Little (1991) and Schenker and Taylor (1996). However, it differs from both approaches in that a model for an ordered categorical variable needs to be specified. More specifically, the following steps create one single set of imputations. First, a bootstrap sample of size N_{obs} from the set of cases having complete data on all variables are drawn. This sample is used to estimate the parameters of a regression of z_{n_1} on all other variables. Second, estimated residuals from the complete cases are randomly selected to predict values — given the other variables — to be imputed for the cases having missing data on the variable DS.

The model used in the first step is a linear regression of the ranks of variable DS, z_{n_1}, on all other variables (Imam and Conover, 1983), i.e. 'age' ('AGE'), 'gender' ('GE'), 'stressful life events' ('SLE') at each quarter of a one year interval after discharge and 'having a depressive espisode' within each quarter of that year. Ties are handled by assigning average ranks. In the second step, the estimated regression parameters are then used to calculate the estimated residuals for the complete cases and the conditional means $\hat{z}^*_{n_1}$ and $\hat{z}^*_{n_2}$, where $\hat{z}^*_{n_l} = \hat{\mathrm{E}}(z_{n_l}|\text{all other variables and estimated parameters})$. One way to proceed would be to randomly draw one of the residuals and predict z_{n_2} by the sum of this residual and the conditional mean $\hat{z}^*_{n_2}$ ('residual draw imputation'). However, to adjust for local lack of fit of the regression model used, we use a 'local residual draw imputation' technique (e.g. Heitjan and Little, 1991; Schenker and Taylor, 1996). More precisely, for each case with missing data five complete cases are found that are closest to the conditional mean $\hat{z}^*_{n_2}$ in the sense of smallest values $d = |\hat{z}^*_{n_1} - \hat{z}^*_{n_2}|$. For every case with incomplete data, one of the corresponding five estimated residuals is randomly drawn to create the predicted value $\hat{z}_{n_2}$ by the sum of the conditional mean $\hat{z}^*_{n_2}$ and this residual. After the values (predicted ranks) are created, new ranks are assigned given all N observations. Again, ties are handled by assigning average ranks. Given the completed data set, the estimates according to the mixed approach described in section 3, the pseudo R^2_T or various test statistics can be calculated. The above steps are repeated M times to create M sets of imputation.

Note that no distributional assumption concerning the model for creating the values to impute are made. On the other hand, the imputation method used is restricted to the case of missingness at random (Rubin, 1987). This assumption, however, seems not to be violated in the present case as will become apparent in section 6.

Inference from multiply imputed data is straightforward (see Rubin, 1987). In particular, let $\hat{\xi}_m$ be a scalar estimate and $\widehat{\mathrm{Var}}_m$ its estimated variance for the mth completed data set. Note, that the estimates using the mixed approach are asymptotically normal. The final estimate is $\hat{\xi} = M^{-1}\sum_m \hat{\xi}_m$ with estimated variance

$$\widehat{\mathrm{Var}} = \overline{\mathrm{Var}} + (1 + M^{-1})\mathrm{B},$$

where $\overline{\mathrm{Var}} = M^{-1}\sum_m \widehat{\mathrm{Var}}_m$ is the average variance within the completed data sets and $\mathrm{B} = (M-1)^{-1}\sum_m(\hat{\xi}_m - \hat{\xi})^2$ is the between imputation variance. Tests of the parameter are based on a t reference distribution with degrees of freedom $v = (M-1)(1+r^{-1})^2$, where $r = (1+M^{-1})\overline{\mathrm{Var}}^{-1}\mathrm{B}$.

If instead of a scalar quantity a k-dimensional estimate $\hat{\xi}$ is of interest with $\hat{\xi}_m$ an estimate and $\widehat{\mathrm{Cov}}_m$ its estimated variance for the mth completed data set, then for the hypothesis $\mathrm{H}_0 : \xi = \xi_0$ Rubin (1987) (see also Rubin and Schenker, 1991) proposes the test statistic

$$\mathrm{d} = [(1+r)k]^{-1}(\hat{\xi} - \xi_0)'\,\overline{\mathrm{Cov}}^{-1}(\hat{\xi} - \xi_0),$$

where $\hat{\xi}$ and $\overline{\mathrm{Cov}}$ are calculated as above and r is generalized to

$$r = (1 + M^{-1})\mathrm{tr}(\mathrm{B}\,\overline{\mathrm{Cov}}^{-1})/k.$$

Tests are based on a $\mathrm{F}_{k,w}$ reference distribution with k and w degrees of freedom, where for $k(M-1) > 4$, w is given by $w = 4 + [(M-1)k - 4](1 + a/r)^2$ and $a = \{1 - 2/[k(M-1)]\}$. If $k(M-1) \leq 4$, then $w = (k+1)v/2$.

5 A Simulation Study

To obtain an idea on how efficient the GEPSE estimator is relative to the estimator proposed by Qu et al. (1992) and Qu et al. (1994) and to a ML estimator, a simple simulation study was conducted where all programs were written using the 'interactive matrix language' (IML) included in the SAS system (SAS Institute Inc., 1989).

Samples were generated according to a panel model with $T = 4$ observations within each block. For each of the $s = 300$ replications $N = 500$ blocks were generated. Two covariates were generated, one uniformly distributed variate varying over all NT observations and one time invariant ordered categorical variate. The latter variate was generated having 30 different values with probabilities equal to the relative frequencies found in the data set described in section 6 for the cases with complete data. Instead of using the ordered categorical variate, a variate the values of which were the ranks of the values of the ordered categorical variate was created (denoted as z). The 'true' values of the parameters were $\beta_{\mathrm{c}} = -1.5$ for the constant term, $\beta_{\mathrm{u}} = -.5$ and $\beta_{\mathrm{r}} = .005$ weighting the uniformly distributed variate and the variate the values of which are the ranks of the ordered categorical variate, respectively. The error terms were generated as standard normally distributed variates according to an equicorrelation structure with $\rho_{t,t'} = \vartheta = .2$ for all t, t' (Model I) and $\rho_{t,t'} = \vartheta = .8$ for all t, t' (Model II), respectively. The 'observable' binary responses were generated according to the threshold model as described in section 2.

After a complete data set was generated, approximately 30% of the values of z were discarded according to the following mechanisms. A continuous variate, r_n^* was generated which was correlated with the uniformly distributed, time invariant variates $x_{n1}, \ldots, x_{n4}$ with correlations .5. A binary variable, r_n, indicating whether the value of z_n is to be deleted was generated according to the rule $r_n = 1$ if $r_n^* > c$ and 0 else. The threshold c was choosen to lead to approximately 30% missing values of the variate z. The two missing mechanisms differed in that in the first case a zero correlation (Mis_1) and in the second a .25 correlation between the two covariates was generated (Mis_2).

The combination of two different values for ϑ and two different missing mechanisms leads to four different situations. Three different estimators were calculated given each of the different situations. The estimator proposed by Qu et al. (1992) and Qu et al. (1994) will be denoted as GEE estimator. The estimator using the mixed approach described in section 3 will be denoted as GEPSE estimator. Both estimators are calculated under the assumption of an equicorrelation structure in the correlation matrix of the latent errors. As ML estimator the maximum likelihood estimator of a simple random effects probit model (e.g. Butler and Moffit, 1982), where it is assumed that $v_{nt} = \pi_n + \epsilon_{nt}$, $\pi_n \sim N(0, \sigma_\pi^2)$, $\epsilon_{nt} \sim N(0, \sigma_\epsilon^2)$ and $\mathrm{E}(\pi_n \epsilon_{nt}) = 0$, restricting the error variance, $\sigma^2 = \sigma_\pi^2 + \sigma_\epsilon^2$, to unity was calculated. The ML estimator of this model can approximately be calculated using Gauss-Hermite quadrature. However, to keep the approximation error under a predefined level, a sufficient number of points for the approximative evaluation of the integrals in the log likelihood function and its derivatives has to be used. The necessary number of evaluation points mainly depends on the value of ϑ. The higher the value of ϑ, the larger the number of points needed and vice versa. Therefore, if $\vartheta = .8$ then 64 evaluation points were used for the estimation results to be stable up to four significant digits. If $\vartheta = .2$, only 34 evaluation points were used.

For every completed data set not only the regression and correlation structure parameter estimates but also pseudo R_T^2 and, since the underlying responses are available, R_T^2 was calculated. For a given simulated data set $M = 4$ sets of imputations were generated. To predict the values to be imputed for each case with missing data, five neighbours were used to randomly draw an estimated residual from. The 'final' estimates, $\hat{\beta}_{\mathrm{c}}, \hat{\beta}_{\mathrm{u}}, \hat{\beta}_{\mathrm{r}}$, using the GEE, GEPSE or ML approach, as well as their estimated variances were calculated as described in section 4. The 'final' estimates $\hat{\vartheta}$ as well as the values of pseudo R_T^2 and R_T^2 were calculated using Fisher's Z-transformation. Corresponding transformations were used to calculate the estimate of the variances of $\hat{\vartheta}$.

To compare the results, the following measures were used: (1) the arithmetic mean of the 'final' estimates over the $s = 300$ replications[3] (M), (2) the estimated standard deviation defined as $\widehat{\mathrm{SD}} = (s^{-1} \sum_{r=1}^{s} \widehat{\mathrm{Var}}(\hat{\theta}_{kr}))^{1/2}$, where $\widehat{\mathrm{Var}}(\hat{\theta}_{kr})$ is the estimated asymptotic variance of the kth element of $\hat{\theta}_r$ $(r = 1, \ldots, s)$, (3) the root mean squared error (RMSE) of the estimates, defined as $\mathrm{RMSE} = (s^{-1} \sum_{r=1}^{s} (\hat{\theta}_{kr} - \theta_k)^2)^{1/2}$ and (4) the proportion of rejections (REJ) at the 5% level of significance of the null hypothesis that the parameter is identical to the 'true' value against a two-sided alternative.

To save space, only the results for Model II are given. For $\vartheta = .2$ (Model I) the differences between GEE, GEPSE and ML estimators for both, Mis_1 and Mis_2, were negligible with respect to the measures defined above. The picture becomes quite different for $\vartheta = .8$ (Model II, see Table 1). If the regression parameters are considered, there is virtually no difference between the GEE and GEPSE estimators concerning the measures M, $\widehat{\mathrm{SD}}$ and RMSE. The difference between these two approaches and the ML approach is only small, with the ML estimator being slightly more efficient in terms of smaller $\widehat{\mathrm{SD}}$ and RMSE. However, the GEE and GEPSE approaches clearly differ with respect to the correlation structure parameter: The GEPSE approach leads to a correlation structure parameter estimator which has smaller $\widehat{\mathrm{SD}}$ and RMSE under both missing mechanisms. Again, the difference between the GEPSE and the ML estimator is only small. To summarize, considering the measures $\widehat{\mathrm{SD}}$ and RMSE for all parameters, the most efficient estimator is the ML estimator followed by the GEPSE estimator. The GEE estimator is the most inefficient estimator, if $\vartheta = .8$.

Insert Table 1 about here

Since all experiments were conducted with $s = 300$ replications, the critical values for a test of the hypothesis of the proportions of rejections being 0.05 are approximately 0.05 ± 0.025 $(\alpha = 0.05)$. The statistic REJ lies outside this interval in five cases, however close to the bounds and in an non-systematic way (see Table 1).

[3]The arithmetic means of R_T^2 and pseudo R_T^2 were calculated using Fishers' Z-transformation.

Note that the above results are valid for both missing mechanisms, although Mis_2 leads to missings which are not missing at random, since the variate r is no more independent of the variate z. For the example considered, it may be concluded that at least in the case of a slight violation of the missing at random assumption all three estimation approaches still lead to satisfactory results.

The arithmetic means of R_T^2 are .51 and .48 for Mis_1 and Mis_2, respectively. The arithmetic means of pseudo R_T^2 are .51 and .48 for Mis_1 and Mis_2, respectively. These arithmetic means are the same for all three estimation approaches. Clearly, as in the univariate case, the arithmetic means of the values of pseudo R_T^2 are very close to those of R_T^2. This result also holds for Model I, i.e. if $\vartheta = .2$.

It should be noted, that the same general results were obtained if the same missing mechanisms and models as above, except that $s = 500$, $N = 250$ and $\beta_r = .01$, were used.

6 Example

In this section the GEPSE approach described in section 3 together with the multiple imputation method described in section 4 is illustrated using a sample of depressed inpatients with a major depression according to DSM-III-R.

The dataset[4] consisted of 139 individuals, 93 females and 46 males, with mean age of 50.4 (SD = 14.7). The data were collected at two points in time, i.e. at discharge from the psychiatric hospital and one year after discharge. The year after discharge was divided into four 3-month intervals. At the end of this year subjects were asked to remember relevant information for the time varying variates using different clues in time as e.g. birthdays or holidays.

The following variates were assumed to have an impact on the probability of having a depressive episode within each of the four 3-month intervals after discharge: age ('AGE') and gender ('GE'; female: 0, male: 1) of the patients, the rank of their depression score at discharge ('DS', Beck Depression Inventory) as time invariant covariates and whether or not stressful life events were experienced ('SLE') within each of the four 3-month intervals as a time varying covariate. A depressive episode was defined by fulfilling the operational criteria and the number of symptoms required for a major depression according to DSM-III-R. Several other variables occasionally considered to be relevant in prediction research for the course of depression (e.g. number of previous episodes, dysthymia; see e.g. Belsher and Costello, 1988; Keller, 1990) were omitted since preliminary analysis failed to confirm substantial relations to relapse. To control for time effects three dummy variables were also included in the model (TIME2 – TIME4), where the first time interval served as reference category.

The values of the variable DS were observed only for $N = 107$ patients (73 females and 34 males with mean age of 49.9 and SD = 14.3). However, this

[4]The data set is available upon request from the first author.

was due to organizational reasons and did not depend upon the values of DS. Therefore, it is reasonable to assume that the values are missing at random in the sense of Rubin (1987).

Beside the effects of the covariates, the structure of the correlation matrix as well as the values of the corresponding correlations between the underlying 'depressivity' given the covariates were of interest. Two different correlation structures are considered. An equicorrelation structure would point to persisting factors not controlled for in the systematic part of the regression model (e.g. a disposition or persisting environmental variables) influencing the probability of being in one of the two states, i.e. having a depressive episode vs. having no depressive episode, at different points in time. On the other hand, an autocorrelation-like structure would point to short-time effects of factors — again not controlled for in the systematic part — leading to a higher probability of being in different states at different points in time. To distinguish between the contributions of long-time vs. short-time effects, a Toeplitz correlation matrix was modeled, i.e. $\mathrm{f}(\vartheta) = (\vartheta_1, \vartheta_2, \vartheta_1, \ldots, \vartheta_{T-1}, \ldots, \vartheta_1)'$.

The regression and correlation structure parameters of this model were estimated using the mixed approach and $M = 4$ completed data sets. For every case with missing data, the estimated residual used to predict the value to be imputed was randomly drawn from five neighbours. Beside the estimates of the parameters and their variances, the pseudo R^2_{Tm} and an estimate defined as $\hat{\zeta}_m = \mathrm{C}\,\hat{\vartheta}_m$, where $\mathrm{C} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}$, together with its variance estimate was calulated for every completed data set. The latter estimates were used to calculate the test statistic d and test the hypothesis $\mathrm{H}_0 : (\vartheta_1 = \vartheta_3 \wedge \vartheta_2 = \vartheta_3)$. The 'final' estimates were calculated as described in section 4. The estimation results of this model are presented in Table 2.

Insert Table 2 about here

From the results in Table 2 it may be concluded, that both the depressive score at discharge and experiencing stressful life events seem to have an effect ($\alpha = .05$) on the probability of having an episode of major depression within a one-year interval after discharge: The higher the depressive score at discharge, the higher the probability of having a depressive episode. Experiencing stressful life events also leads to a higher probability of having a relapse. Corresponding effects cannot be shown for the other covariates. The estimates $\hat{\vartheta}_1$, $\hat{\vartheta}_2$ and $\hat{\vartheta}_3$ point to an autocorrelation-like association structure, i.e. smaller correlations with increasing distances in time. Since the hypothsis $\mathrm{H}_0 : (\vartheta_1 = \vartheta_3 \wedge \vartheta_2 = \vartheta_3)$ can be rejected at the 5% level and ϑ_3 cannot be shown to be significantly different from zero whereas ϑ_1 and ϑ_2 are significantly different from zero at the 5% level, it may be concluded that persistent factors not accounted for in the systematic part of the regression model cannot be shown to be important with respect to the probability of having or not having an episode of major depression.

The value of pseudo R_T^2 is rather low. Although the impact of two covariates can assumed to be significant, the proportion of variance 'explained' by the systematic part of the model is small. Clearly, more work is needed to identify additional variables having a significant impact on the probability of experiencing a relapse into a major depression. The search for those variables is assisted by the results concerning the correlation structure, since they suggest to search for variables having only temporarily limited effects. On the other hand, it should be noted that it is questionable whether a relatively small set of variables is responsible for having a relapse. Rather many variables are expected to have only moderate effects, maybe interacting in a complicated way (c.f. Kendler, Kessler, Neale, Heath & Eaves, 1993). This, however, leads to the necessity of considerable larger sample sizes which is prohibitive in many applications. Therefore, the example shows that it may not only be important to focus on exogenous, observable variates but also to account for variables not explicitely considered in the systematic part of the model which then enter into the error term of the model. Furthermore, of course, modeling the correlation structure carefully simply leads to more efficient estimators of the regression parameters.

7 Discussion

The approach proposed in section 3 allows the estimation of multivariate probit models. In the present paper, a model for binary clustered or longitudinal data as a special case of a multivariate probit is considered. Although the results of the simulation study in section 5 should not be overgeneralized, they suggest the GEPSE approach — at least for the models considered — to lead to estimators which are — in finite samples with missing values of an ordered categorical time invariant variable — very efficient relative to the ML estimator and for high 'true' correlations to more efficient estimators relative to the GEE estimator proposed by Qu et al. (1992) and Qu et al. (1994). Furthermore, the proposed multiple imputation technique using a regression of the ranks of the values of a variate with missing values on all other variabels for those cases having complete data to predict missing values worked well in the simulations. The proposed pseudo R_T^2 was found to be close to the 'true' R_T^2 for the underlying linear model and can therefore be recommended for applications. This result mirrors corresponding results in the univariate case (e.g. Veall & Zimmermann, 1992; Veall & Zimmermann, 1996; Windmeijer, 1995). Clearly, more systematic simulations are necessary to assess the properties of the GEPSE estimator relative to alternative estimators using the proposed multiple imputation technique under different standard and non-standard conditions.

However, the above results suggest, that if one is interested not only in the regression parameters but also in the correlations of the errors of the underlying model or in functions thereof, than the GEPSE approach is recommended if no ML estimator is available.

Although a non-orthogonal estimation of the two sets of parameters would be possible in the same way as described in Zhao and Prentice (1990) for the GEE estimators, the efficiency gain can be expected to be only negligible. Furthermore, the robustness property of the regression parameters with respect to misspecification of the association structure would be lost.

Several generalizations to the mixed approach are possible. For example, in a slightly more general framework, the regression parameter estimates are not restricted to be identical over different observations within blocks. Furthermore, the model may be extended to handle ordered or unordered categorical responses as well.

In the example presented in Section 6 an AR(1)-like structure was found in the estimated correlation matrix pointing to short-time effects of factors not accounted for in the systematic part of the regression model. However, a structure like this could also arise from variables not accounted for in the systematic part of the model being autocorrelated and not independent from the covariates included in the model. In this case an assumption made in Section 2 would be violated and the regression estimator may not be consistent any more. As this cannot be completely ruled out in the example presented, as well as in many other applications, it is not yet clear how severe the effects on the properties of the estimators are if one or more of the different assumptions are violated. This is a point that clearly needs more careful investigation.

Appendix

Let $\theta = (\beta', \vartheta')'$, $\mathrm{u}_1(\beta,\vartheta) = \sum_n \mathrm{A}_n'\Omega_n^{-1}\mathrm{e}_n$ and $\mathrm{u}_2(\beta,\vartheta) = \frac{\partial\rho}{\partial\vartheta}\sum_n \mathrm{B}_n'\mathrm{W}_n^{-1}\mathrm{v}_n$, then, using a Taylor expansion, $\sqrt{N}(\hat{\theta}-\theta_0)$ can — under some regularity conditions — be approximated by

$$-\begin{bmatrix} N^{-1}\frac{\partial\,\mathrm{u}_1(\beta,\vartheta)}{\partial\,\beta} & N^{-1}\left(\frac{\partial\,\mathrm{u}_1(\beta,\vartheta)}{\partial\,\vartheta}\right)' \\ N^{-1}\left(\frac{\partial\,\mathrm{u}_2(\beta,\vartheta)}{\partial\,\beta}\right)' & N^{-1}\frac{\partial\,\mathrm{u}_2(\beta,\vartheta)}{\partial\,\vartheta} \end{bmatrix}^{-1}_{\theta=\theta_0} \begin{bmatrix} N^{-1/2}\mathrm{u}_1(\beta,\vartheta) \\ N^{-1/2}\mathrm{u}_2(\beta,\vartheta) \end{bmatrix}_{\theta=\theta_0}. \quad \text{(A.1)}$$

It can then be shown, that the functions $(N^{-1/2}\mathrm{u}_1'(\beta,\vartheta),\ N^{-1/2}\mathrm{u}_2'(\beta,\vartheta))'_{\theta=\theta_0}$ have an asymptotic normal distribution as $N\to\infty$ with mean zero and covariance matrix

$$\lim_{N\to\infty} N^{-1} \begin{bmatrix} \sum_n \mathrm{A}_n'\Omega_n^{-1}\mathrm{Cov}(y_n)\Omega_n^{-1}\mathrm{A}_n & \left(\sum_n \mathrm{A}_n'\Omega_n^{-1}\mathrm{E}(\mathrm{e}_n\mathrm{v}_n'\mathrm{W}_n^{-1})\mathrm{B}_n\right)\frac{\partial\rho}{\partial\vartheta'} \\ \frac{\partial\rho}{\partial\vartheta}\sum_n \mathrm{B}_n'\mathrm{E}(\mathrm{W}_n^{-1}\mathrm{v}_n\mathrm{e}_n')\Omega_n^{-1}\mathrm{A}_n & \frac{\partial\rho}{\partial\vartheta}\left(\sum_n \mathrm{B}_n'\mathrm{E}(\mathrm{W}_n^{-1}\mathrm{v}_n\mathrm{v}_n'\mathrm{W}_n^{-1})\mathrm{B}_n\right)\frac{\partial\rho}{\partial\vartheta'} \end{bmatrix}_{\theta=\theta_0}. \quad \text{(A.2)}$$

Again, under mild regularity conditions, it can be shown that as $N\to\infty$

$$\begin{aligned} N^{-1}\frac{\partial\,\mathrm{u}_1(\beta,\vartheta)}{\partial\,\beta} &\overset{p}{\longrightarrow} -N^{-1}\left(\sum_n \mathrm{A}_n'\Omega_n^{-1}\mathrm{A}_n\right)_{\theta=\theta_0}, \\ N^{-1}\frac{\partial\,\mathrm{u}_1(\beta,\vartheta)}{\partial\,\vartheta} &\overset{p}{\longrightarrow} 0, \\ N^{-1}\frac{\partial\,\mathrm{u}_2(\beta,\vartheta)}{\partial\,\beta} &\overset{p}{\longrightarrow} N^{-1}\left(\sum_n \frac{\partial^2\,l_n(\vartheta,\beta)}{\partial\,\beta\,\partial\,\vartheta'}\right)_{\theta=\theta_0} \end{aligned}$$

and

$$N^{-1}\frac{\partial\,\mathrm{u}_2(\beta,\vartheta)}{\partial\,\vartheta} \overset{p}{\longrightarrow} N^{-1}\left(\sum_n \frac{\partial^2\,l_n(\vartheta)}{\partial\,\vartheta\,\partial\,\vartheta'}\right)_{\theta=\theta_0}.$$

Therefore, the matrix in (A.1) converges as $N\to\infty$ to

$$\lim_{N\to\infty} N^{-1}\begin{bmatrix} -\sum_n \mathrm{A}_n'\Omega_n^{-1}\mathrm{A}_n & 0 \\ \sum_n \frac{\partial^2\,l_n(\vartheta,\beta)}{\partial\,\vartheta\,\partial\,\beta'} & \sum_n \frac{\partial^2\,l_n(\vartheta)}{\partial\,\vartheta\,\partial\,\vartheta'} \end{bmatrix}_{\theta=\theta_0}. \quad \text{(A.3)}$$

Combining (A.3) with (A.2), inserting estimates $\hat{\theta}$ for θ and estimating $\mathrm{Cov}(y_n)$, $\mathrm{E}(\mathrm{e}_n\mathrm{v}_n'\mathrm{W}_n^{-1})$, $\mathrm{E}(\mathrm{W}_n^{-1}\mathrm{v}_n\mathrm{e}_n')$ and $\mathrm{E}(\mathrm{W}_n^{-1}\mathrm{v}_n\mathrm{v}_n'\mathrm{W}_n^{-1})$ by $\mathrm{e}_n\mathrm{e}_n'$, $\mathrm{e}_n\mathrm{v}_n'\mathrm{W}_n^{-1}$, $\mathrm{W}_n^{-1}\mathrm{v}_n\mathrm{e}_n'$ and $\mathrm{W}_n^{-1}\mathrm{v}_n\mathrm{v}_n'\mathrm{W}_n^{-1}$, respectively, leads to the covariance matrix estimator described in Section 3.

References

Anderson, D.A. & Aitkin, M. (1985). Variance component models with binary responses: Interviewer Variability. *Journal of the Royal Statistical Society, Ser. B, 47(2),* 203–210.

Ashford, J.R. & Sowden, R.R. (1970). Multi-variate probit analysis. *Biometrics, 26,* 535–546.

Avery, R.B., Hansen, L.P. & Hotz, V.J. (1983). Multiperiod probit models and orthogonality condition estimation. *International Economic Review, 24(1)*, 21–35.

Belsher, G. & Costello, C.G. (1988). Relapse after recovery from unipolar depression: A critical review. *Psychological Bulletin, 104*, 84–96.

Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46(4)*, 443–459.

Bock, R.D. & Gibbons, R.D. (1996). High-dimensional multivariate probit analysis. *Biometrics, 52,* 1183–1194.

Bock, R.D. & Lieberman, M. (1970). Fitting a reponse model for n dichotomously scored items. *Psychometrika, 35(2),* 179–197.

Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88(241),* 9–25.

Butler, J.S. & Moffit, R. (1982). Notes and comments: A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica, 50(3)*, 761–764.

Conaway, M.R. (1989). Analysis of repeated categorical measurements with conditional likelihood methods. *Journal of the Amercian Statistical Association, 84(405),* 53–62.

Gibbons, R.D., Hedeker, D., Charles, S.C. & Frisch, P. (1994). A random-effects probit model for predicting medical malpractice claims. *Journal of the American Statistical Association, 89(427),* 760–767.

Gourieroux, C., Monfort, A. & Trognon, A. (1984). Estimation and test in probit models with serial correlation. In J.P. Florens, M. Mouchart, J.P. Raoult & L. Simar (Eds.), *Alternative approaches to time series analysis, Proceedings of the 3rd Franco-Belgian Meeting of Statisticians, november*

1982 (pp. 169–209). Bruxelles: Publications des Facultès universitaires Saint-Louis, Boulevard de Jardin botanique 43.

Heckman J.J. (1981). Statistical models for discrete panel data. In C.F. Manski, & D. McFadden, (eds.), *Structural analysis of discrete data with econometric applications* (pp. 114–178). Cambridge: The MIT Press.

Hedeker, D. & Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics, 50,* 933-944.

Heitjan, D.F. & Little, R.J.A. (1991). Multiple Imputation for the fatal accident reporting system. *Applied Statistics, 40(1),* 13–29.

Hooper, J.W. (1959). Simultaneous equations and canonical correlation theory. *Econometrica, 27,* 245–256.

Im, S. & Gianola, D. (1988). Mixed models for binomial data with an application to lamb mortality. *Applied Statistics, 37(2),* 196–204.

Imam, R.L. & Conover, W.J. (1983). *Modern business statistics.* New York: John Wiley.

Keller, M.B. (1990). Diagnostic and course-of-illness variables pertinent to refractory depression. In A. Tasman, S.M. Goldfinger & C.A. Kaufmann (Eds.), *American Psychiatric Press Review of Psychiatry, Volume 9* (pp. 10–32). Washington: American Psychiatric Press.

Kendler, K.S., Kessler, R.C., Neale, M.C., Heath, A.C. & Eaves, L.J. (1993). The prediction of major depression in women: Toward an integrated etiologic model. *American Journal of Psychiatry, 150,* 1139–1148.

Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Ser. B, 58(4),* 619–678.

Liang, K.-Y. & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73(1)*, 13–22.

Liang, K.Y., Zeger, S. & Qaqish, B. (1992). Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, B, 54(1),* 3–40.

Little, R.J.A. (1992). Regression with missings X's: A review. *Journal of the American Statistical Association, 87(420),* 1227–1237.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: John Wiley & Sons.

McGilchrist, C.A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society, B, 56(1),* 61–69.

McKelvey, R.D. & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4,* 103–120.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49(1)*, 115–132.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. – VIII. On the inheritance of characters not capable of exact quantitative measurement. *Philosophical Transactions Of The Royal Society Of London, Series A, 195*, 79–150.

Pendergast, J.F., Gange, S.J., Newton, M.A., Lindstrom, M.J., Palta, M. & Fisher, M.R. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review, 64(1),* 89–118.

Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics, 44*, 1033–1048.

Qu, Y., Williams, G.W., Beck, G.J. & Medendorp, S.V. (1992). Latent variable models for clustered dichotomous data with multiple subclusters. *Biometrics, 48*, 1095–1102.

Qu, Y., Piedmonte, M.R. & Williams, G.W. (1994). Small sample validity of latent variable models for correlated binary data. *Communications in Statistics — Simulation and Computation, 23(1)*, 243–269.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91(434),* 473–489.

Rubin, D.B. & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association, 81(394),* 366–374.

Rubin, D.B. & Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine, 10,* 585–598.

SAS Institute Inc. (1989). *SAS/IML software: Usage and reference, version 6.* Cary, NC: SAS Institute.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika, 78(4),* 719–727.

Schenker, N. & Taylor, J.M.G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis, 22,* 425–446.

Schepers, A., Arminger, G. & Küsters, U. (1991). The analysis of non-metric endogenous variables in latent variable models: The MECOSA approach. In J. Gruber (Ed.), *Econometric Decision Models: New Methods of Modelling and Applications, Lecture Notes in Economics and Mathematical Systems: 366* (pp. 459–472). Berlin: Springer.

Spiess, M. & Hamerle, A. (1995). *Regression models with correlated binary response variables: A comparison of different methods in finite samples* (Discussion Paper No. 10). München: SFB 386 der Ludwig-Maximilians-Universität-München. (http://www.stat.uni-muenchen.de/sfb386/publikation.html)

Stiratelli, R., Laird, N. & Ware, J.H. (1984). Random-effects models for serial observations with binary responses. *Biometrics, 40,* 961–971.

Veall, M.R. & Zimmermann, K.F. (1992). Pseudo-R^2's in the ordinal probit model. *Journal of Mathematical Sociology, 16,* 333–342.

Veall, M.R. & Zimmermann, K.F. (1996). Pseudo-R^2 measures for some common limited dependent variable models. *Journal of Economic Surveys, 10(3),* 241–259.

Windmeijer, F.A.G. (1995). Goodness-of-fit measures in binary choice models. *Econometric Reviews, 14(1),* 101–116.

Wong, G.Y. & Mason, W.M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association, 80(391),* 513–524.

Zhao, L.P. & Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika, 77,* 642–648.

Table 1: *Mean (M), estimated standard deviation ($\widehat{SD}$) root mean squared error (RMSE) and proportion of rejections (REJ) for Model II with $N = 500$, $T = 4$, 30% missing values, $\hat{\beta}_c = -1.5$, $\hat{\beta}_u = -.5$ and $\hat{\beta}_r = .005$ over $s = 300$ replications*

M $\widehat{\text{SD}}$ RMSE REJ	GEE	Mis_1 GEPSE	ML	GEE	Mis_2 GEPSE	ML
$\hat{\beta}_\text{c}$	-1.479	-1.479	-1.474	-1.495	-1.495	-1.491
	.1413	.1413	.1405	.1377	.1377	.1371
	.1223	.1222	.1226	.1206	.1208	.1199
	.03	.027	.037	.023	.023	.027
$\hat{\beta}_\text{u}$	-.5003	-.5004	-.5000	-.5045	-.5045	-.5041
	.0357	.0357	.0352	.0364	.0364	.0361
	.0334	.0334	.0334	.0332	.0332	.0331
	.05	.05	.047	.037	.037	.03
$\hat{\beta}_\text{r}$	.0049	.0049	.0049	.0050	.0050	.0050
	.0005	.0005	.0005	.0005	.0005	.0005
	.0004	.0004	.0004	.0004	.0004	.0004
	.04	.04	.047	.027	.027	.023
$\hat{\vartheta}$	.7965	.8001	.8003	.7970	.7997	.7994
	.0354	.0287	.0282	.0336	.0277	.0275
	.0317	.0252	.0242	.0302	.0247	.0245
	.02	.04	.04	.037	.027	.02

Table 2: *Estimates, estimated standard deviations ($\widehat{SD}$), degrees of freedom (v) and t-values using $M = 4$ completed data sets*

	Estimate	$\widehat{\text{SD}}$	v	$t - value$
Intercept	-1.242	0.319	322	-3.89
AGE	-0.004	0.005	> 500	-0.84
GE	0.130	0.185	> 500	0.70
DS	0.011	0.002	47	4.41
SLE	0.301	0.124	> 500	2.42
TIME2	0.178	0.102	> 500	1.75
TIME3	0.147	0.139	> 500	1.06
TIME4	0.096	0.153	> 500	0.63
ϑ_1	0.794	0.046	> 500	19.49
ϑ_2	0.432	0.110	> 500	3.88
ϑ_3	0.228	0.156	> 500	1.51

pseudo $\mathrm{R}^2 = .15$

d $= 10.98$ with $k = 2$ and $w = 1281$